Salient in space - salient in time: fixation probability predicts fixation duration during natural scene viewing

Salient in space, salient in time: Fixation probability predicts fixation duration during natural scene viewing

Wolfgang Einhäuser

Institute of Physics, Technische Universität Chemnitz, Chemnitz, Germany
Neurophysics, Philipps-University Marburg, Marburg, Germany

Antje Nuthmann

Psychology Department, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh, UK

During natural scene viewing, humans typically attend and fixate selected locations for about 200–400 ms. Two variables characterize such "overt" attention: the probability of a location being fixated, and the fixation's duration. Both variables have been widely researched, but little is known about their relation. We use a two-step approach to investigate the relation between fixation probability and duration. In the first step, we use a large corpus of fixation data. We demonstrate that fixation probability (empirical salience) predicts fixation duration across different observers and tasks. Linear mixed-effects modeling shows that this relation is explained neither by joint dependencies on simple image features (luminance, contrast, edge density) nor by spatial biases (central bias). In the second step, we experimentally manipulate some of these features. We find that fixation probability from the corpus data still predicts fixation duration for this new set of experimental data. This holds even if stimuli are deprived of low-level images features, as long as higher level scene structure remains intact. Together, this shows a robust relation between fixation duration and probability, which does not depend on simple image features. Moreover, the study exemplifies the combination of empirical research on a large corpus of data with targeted experimental manipulations.

Introduction

Scrutinizing long fixations observers made when viewing a picture of a painting, Buswell hypothesized in 1935 that "the main centers of interest, as judged by number of fixations, also receive the fixations which are longest in duration" (p. 90). He thereby linked two fundamental aspects of eye guidance in scene viewing: fixation number or fixation probability (*Where* do the eyes preferentially fixate?) and fixation duration (*When* do the eyes proceed to the next location?). In the eight decades since Buswell's seminal study, both aspects of gaze guidance have received considerable research interest, but they have been accounted for separately. This when/where separation goes back to an early suggestion of distinct oculomotor control circuits (van Gisbergen, Gielen, Cox, Bruijns, & Kleine Schaars, 1981). Accordingly, Findlay and Walker (1999) proposed an influential qualitative model of oculomotor control that completely separated the when and where systems. Models of eye-movement control in reading have also adopted this separation (Engbert, Nuthmann, Richter, & Kliegl, 2005; Reichle, Rayner, & Pollatsek, 2003). Empirical and computational research on real-world scene viewing has focused on the where decision to a great extent. The widely used saliency map (Itti, Koch, & Niebur, 1998) and its recent variants have had some success in predicting fixation probability (for a review, see Borji, Sihite, & Itti, 2013a; but see Tatler, Hayhoe, Land, & Ballard, 2011). However, the original implementation of the saliency map failed to achieve realistic fixation durations: Using biophysically realistic time constants, the dwell times of the "focus-of-attention" were too low to be interpreted as fixation durations (Itti et al., 1998); conversely, when model parameters were constrained by search time, the model arrived at unrealistically long fixation durations (Itti & Koch, 2000). Following this lead, more recent developments of salience-type models, which have improved the predictive power for fixation probability (Bruce & Tsotsos, 2009; Erdem & Erdem, 2013; Garcia-Diaz, Leborán, Fdez-Vidal, & Pardo, 2012; Harel, Koch, &

Citation: Einhäuser, W., & Nuthmann, A. (2016). Salient in space, salient in time: Fixation probability predicts fixation duration during natural scene viewing. *Journal of Vision, 16*(11):13, 1–17, doi:10.1167/16.11.13.

doi: 10.1167/16.11.13 Received March 14, 2016; published September 14, 2016 ISSN 1534-7362

Perona, 2007; Lin & Lin, 2014; Xu, Jiang, Wang, Kankanhalli, & Zhao, 2014; Zhang, Tong, Marks, Shan, & Cottrell, 2008), have consistently ignored the issue of fixation duration. Conversely, the CRISP model of fixation durations in scene viewing (Nuthmann & Henderson, 2012; Nuthmann, Smith, Engbert, & Henderson, 2010) models the control of fixation durations without taking fixation locations into account. To summarize, models of eye guidance in natural scene viewing have addressed *either* the where *or* the when decision, but not both.

In addition to proposing a link between the *when* and the *where*, Buswell's statement implicates that it is some form of high-level "interestingness" that guides fixation duration and probability alike. At first glance, this appears in contrast to most contemporary computational approaches, which model fixation probability as consequence of low-level features. However, key proponents of salience models have argued that these models approximate "interesting objects" (Elazary & Itti, 2008) or "objects that stand out" (Borji, Sihite, & Itti, 2013b). In turn, scene regions that are "interesting" (Masciocchi, Mihalas, Parkhurst, & Niebur, 2009), "informative" (Antes, 1974; Mackworth & Morandi, 1967), or "relevant" (Onat, Açık, Schumann, & König, 2014), according to the consensus of other observers, are preferentially fixated. Such human-defined, high-level salience typically outperforms model-defined salience in predicting fixated locations (Koehler, Guo, Zhang, & Eckstein, 2014; Onat et al., 2014). Similarly, the presence of objects overrides low-level salience (Stoll, Thrun, Nuthmann, & Einhäuser, 2015). Despite the success of low-level models, higher level information and scene content play an important role in guiding gaze, in line with Buswell's interestingness assertion. The interaction between low-level salience and higher level content nonetheless necessitates that any putative relation between *when* and *where* has to be controlled for low-level feature effects.

Research in reading and visual search has demonstrated that fixation durations are sensitive to moment-to-moment processing demands (Rayner, 1998), and this generalizes to scene viewing (Nuthmann et al., 2010). A number of scene-viewing studies have investigated the impact of low-level features on fixation durations by virtue of feature modifications to the entire image. As a general finding, any image-wide degradation of low-level features prolongs fixations. For example, fixation durations have been found to increase when the overall luminance of the scene is reduced (Henderson, Nuthmann, & Luke, 2013; Walshe & Nuthmann, 2014), color is removed (Ho-Phuoc, Guyader, Landragin, & Guérin-Dugué, 2012; Nuthmann & Malcolm, 2016), or phase information is removed from the scene ($1/f$ noise; Kaspar & König, 2011; Walshe & Nuthmann, 2015). Moreover, *local* image statistics around fixation also modulate fixation duration (Nuthmann, 2016). Recent fMRI data support the notion that cognitive scene-processing demands control fixation duration in an online ("real-time") fashion (Henderson & Choi, 2015), which is consistent with the CRISP model.

The present study tests Buswell's hypothesis on a relation of fixation probability and duration. We specify linear mixed-effects models (LMMs) to test whether fixation probability at a given location predicts fixation duration at the same location. In the first step, we report new analyses of a large corpus data set, based on 72 observers engaging in three different tasks on 135 images. Fixation probability is computed from the aggregate fixation data from one viewing task; prediction of durations is then done within the same task as well as across tasks, and therefore across observer subsets. In the second step, we control statistically for potentially confounding variables on the hypothesized relation between fixation probability and duration. To do so, we add to the model low-level features (luminance, contrast; Reinagel & Zador, 1999), a mid-level feature (edge density; Baddeley & Tatler, 2006; Mannan, Ruddock, & Wooding, 1996), and a generic spatial bias (central bias; Clarke & Tatler, 2014; Tatler, 2007). In the third step, we control experimentally for such confounders by using the fixation probabilities obtained in the original experiment to predict fixation durations in an independent set of observers who freely viewed feature-degraded versions of the same scenes.

Materials and methods

Experiment 1

Stimuli and data collection

Data for Experiment 1 come from a large corpus of eye movements during scene viewing. The methods of data acquisition have been reported elsewhere (Nuthmann & Henderson, 2010; Pajak & Nuthmann, 2013).[1] In brief, 72 observers each viewed 135 images of natural scenes (43 outdoor and 92 indoor, depicting street views and diverse interior spaces with varying degree of clutter) for 8 s each, while their eye position was recorded with an EyeLink 1000/2K system (SR Research, Ottawa, Canada). For each observer, images were divided into three subsets of 45 images on which three different tasks were performed: search for a verbally defined target, memorization of the image for a subsequent memory test, and an aesthetic-preference judgment. Images had a resolution of 800×600 pixels, and 1° of visual angle corresponded to 32 pixels.

For the present analysis, fixations were excluded if they preceded or succeeded a blink, if they started before image onset or ended after image offset, or if their duration was below 50 ms or above 1000 ms. For

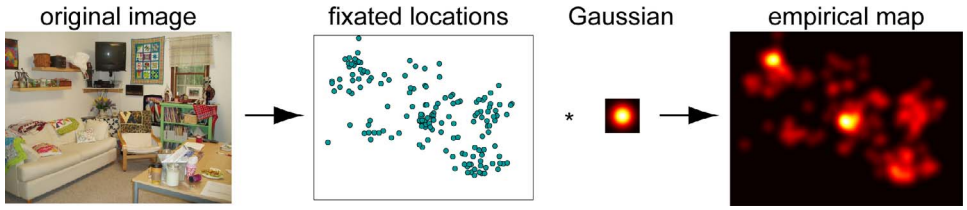

Figure 1. Empirical salience map. Scheme for computing the empirical map for a given image: All fixations for a given task (the predicting task) were pooled and smoothed with a Gaussian kernel.

the search task, fixations made after the search target was found were also excluded. After applying these criteria, 33,162 fixations were included for search, 67,484 for memorization, and 69,847 for the preference-judgment task.

Empirical salience maps

The probability of a location being fixated was quantified by applying the following procedure to each image (Figure 1; cf. Pomplun, Ritter, & Velichkovsky, 1996): First, fixations were pooled over all observers who viewed the image under a given viewing-task instruction (the predicting condition). Second, a Gaussian of 0.5° standard deviation was centered at each fixation. Third, these Gaussians were added. To ease comparison to feature maps, the resulting maps were z-scored—that is, normalized to zero mean and unit standard deviation. Bar this linear scaling, the value at each location corresponds to the probability of the location being fixated. We refer to this value as empirical salience and to the map as an empirical salience map.

Image features

To test whether effects of fixation probability on fixation duration were mediated by low-level image features, we specified models that included local luminance and luminance contrast (low-level features, Figure 2A, B) as well as edge density (mid-level image feature, Figure 2C). For consistency with the definition of the empirical salience maps, all features were defined as weighted mean around the respective location, where the weighting function was a Gaussian of 0.5° standard deviation; mathematical definitions of the features are provided in Appendix A. Eccentricity, defined as the Euclidian distance from the center of the image, was used as additional predictor (Figure 2D). All feature maps were z-normalized to zero mean and unit standard deviation to be on a consistent scale with each other and with the empirical maps.

Following experimental studies on fixation probability (e.g., Baddeley & Tatler, 2006; Einhäuser & König, 2003; Mannan et al., 1996; Reinagel & Zador, 1999; Vincent, Baddeley, Correani, Troscianko, & Leonard, 2009), we chose to probe a number of individual image features rather than reverting to aggregate salience measures that are often used in computational studies (Borji et al., 2013a; Kümmerer, Wallis, & Bethge, 2015). This way, we can also account for dependencies among features (cf. Baddeley & Tatler, 2006; Nuthmann, 2016; Nuthmann & Einhäuser, 2015), which are explicitly modeled by our approach.

Experiment 2

Stimuli

Experiment 2 served to test whether a relation of fixation probability and duration would prevail once various low-level features in these images were modified or removed. Stimuli were based on a subset of 48 images from Experiment 1. When selecting these images, we first excluded the 52 images depicting humans and two images whose original resolution differed from the default. From the remaining images, all 10 outdoor images and the first 38 indoor images were selected for Experiment 2. For all conditions in Experiment 2, color was removed from the images. To this end, images were first converted to grayscale using MATLAB's (The MathWorks, Natick,

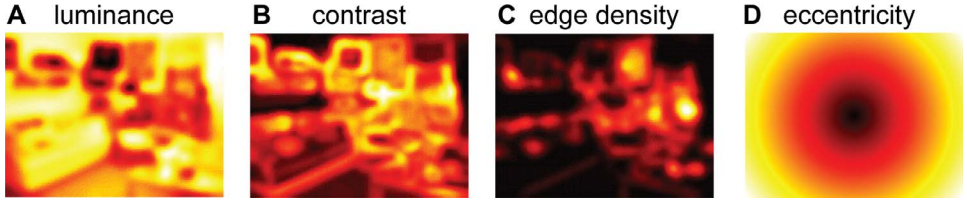

Figure 2. Feature maps for luminance (A), contrast (B), and edge density (C), along with a generic feature map (D) to capture the central bias modeled by eccentricity. All maps are for the example image shown in Figure 1; warmer colors correspond to higher feature values.

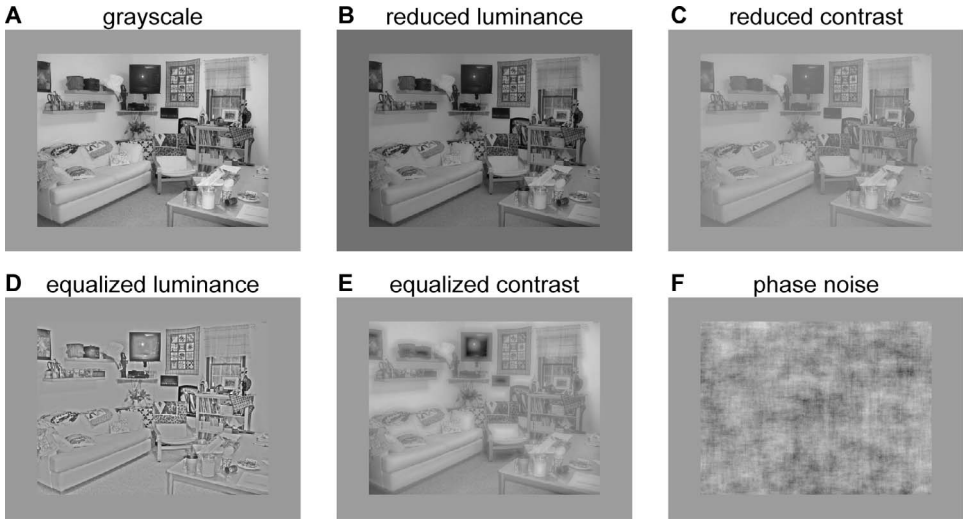

Figure 3. Conditions used in Experiment 2 for the example stimulus displayed in Figure 1.

MA) rgb2gray function. The resulting pixel values were then linearly mapped to a range from 0 to 1. Assuming the unknown camera gamma to be 0.5, which is a typical setting, this intensity was then squared and mapped linearly to the luminance range of the display (0.11 to 66.0 cd/m^2). Based on these grayscale images, five conditions were defined in which low-level features were modified (Figure 3): Condition 1, the grayscale image itself, i.e., the original image deprived of color (Figure 3A); Conditions 2 and 3, images with reduced global luminance and contrast, respectively (Figure 3B, C); and Conditions 4 and 5, images with luminance (Condition 4, Figure 3D) or contrast (Condition 5, Figure 3E) nearly equalized at scales above 0.5°. In the sixth condition, the image's phase spectrum was randomized except for the lowest frequency components. Randomizing the phase of a natural scene is a widely used method to destroy higher order scene structure (e.g., Rainer, Augath, Trinath, & Logothetis, 2001; response by Dakin, Hess, Ledgeway, & Achtman, 2002; Einhäuser et al., 2006; Kayser, Nielsen, & Logothetis, 2006; Rainer, Lee, & Logothetis, 2004; Wichmann, Braun, & Gegenfurtner, 2006). The present modification keeps mean luminance and the luminance autocorrelation unaltered, and in addition preserves information on a very low frequency scale. Higher order structure, including edges, higher order groupings, and objects, is destroyed (Figure 3F). The mathematical details of the modifications are given in Appendix B.

Observers

Twenty-four observers (12 men and 12 women, age range = 20–36 years) participated in Experiment 2. Participants were unaware of the purpose of the experiment. The number of participants was chosen to match the number of participants per task in Experiment 1. Procedures adhered to the Declaration of Helsinki and were approved by the local ethics review board (Ethikkommission FB04, Philipps-Universität Marburg), and all participants gave written informed consent.

Setup

Stimuli were presented on a 19-in. CRT screen (EIZO FlexScan F77S) running at a resolution of 1024 × 768 pixels at a 100-Hz refresh rate. The luminance of the screen was set to range from 0.11 cd/m^2 (black) to 66 cd/m^2 (white). These settings, despite implying a comparably low contrast of 600:1, allowed faithful representation of the low luminance values. The stability and correctness of luminance settings were verified with a Mavo Monitor USB photometer (Gossen, Nuremberg, Germany) at the start and the end of each testing day. Stimuli were presented at their native 800 × 600 resolution centrally in a gray (33.0 cd/m^2) frame. The screen was located 73 cm from the observer, implying that 1° of visual angle corresponded to 32.6 pixels centrally. Eye position was recorded at 1000 Hz with an EyeLink 1000 device. Blink and fixation detection used the EyeLink's built-in algorithm with standard settings (saccade thresholds of 35°/s for velocity and 9500°/s^2 for acceleration).

Procedure

Each observer performed a total of 288 trials, split over six blocks of 48 trials each. In each block, all 48 distinct images were shown once in exactly one of the six conditions. Over the course of the six blocks, each image was shown in each condition once. Per block, each condition was used eight times. The assignment of the 288 distinct stimuli (6 conditions × 48 images) to a given block was balanced across the 24 observers. Within each block, order of stimuli was randomized.

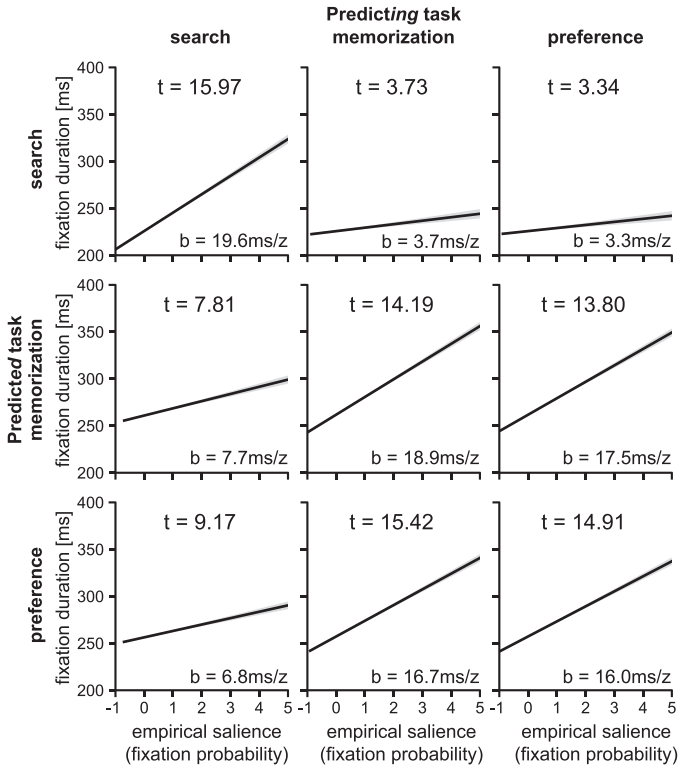

Figure 4. Experiment 1: Prediction of fixation durations by fixation probability. Each panel depicts the partial effect of fixation probability (empirical salience map) on fixation duration. Fixation probabilities (x-axes) were z-scored for model fitting. Regression coefficients (*b*s; that is, the slope of the empirical-map predictor, in ms per z-normalized probability) and *t* values ($t = b/\mathrm{SE}$) are given in each panel. Each row corresponds to a different task that is predicted (predicted condition), each column to a different task from which the empirical salience map is computed (predicting condition).

Observers started each trial by fixating centrally (within 1° of a central black cross) on a gray (33 cd/m^2) screen for at least 300 ms. If they failed to reach stable fixation within 3 s, the eye tracker was recalibrated. After a further 100 ms the stimulus was presented and remained visible for 5 s. Observers were instructed to "study each image carefully" and told that they were "free to move [their] eyes whenever the stimulus is on." After each block, the eye tracker was recalibrated and participants had the opportunity to take a break. For the purpose of a different study, all observers also performed the same task on an unrelated stimulus set in a separate session either directly before (half of observers) or directly after Experiment 2.

Data analysis and models

To test how well empirical salience (i.e., fixation probability) predicted fixation duration in each of the three tasks (the predicted condition), LMMs were used. Each fixation was treated as an individual observation, fixation duration was the dependent variable, and fixation probability was defined as the value of the empirical salience map sampled at the corresponding fixated location. In addition to empirical salience as a fixed effect, each model included random intercepts and random slopes for subjects and items (images). The values of all input variables (empirical salience, luminance, contrast, edge density, eccentricity) were linearly scaled to have zero mean and unit standard deviation (z-scored) prior to computing the LMMs (Schielzeth, 2010). This scaling puts all predictors on a commensurate scale. By definition, the linear scaling does not affect the shape of the distributions of input variables. We deliberately refrained from applying somewhat arbitrary nonlinear transformations, and instead tolerate that some distributions are skewed toward low values.

Simple models

For a first analysis (Figure 4), empirical salience was included as a fixed effect in the LMMs in addition to the intercept. The LMMs had the maximal random-effects structure for subjects and items (Barr, Levy, Scheepers, & Tily, 2013). For subjects there was a random intercept, a random slope for empirical salience, and a correlation parameter for a possible correlation between intercept and slope; the same was true for items (scenes).

Models including image features

For the later analyses (Figures 5 through 8), LMMs additionally included the three image features (luminance, contrast, edge density) and eccentricity as fixed effects. For each of these features, we computed a spatial map (Figure 2). As with the empirical map, for each fixation the values of these maps were sampled at the fixated location. The maximal random-effects structure for these LMMs would require estimating 42 parameters (by subject: random intercept, five random slopes, 15 correlation terms; by item: same as by subject). This maximal random-effects structure is too complex for the information contained in the data, with the result that some of the LMMs did not converge. To reduce model complexity, the correlations between random intercepts and slopes were set to zero (cf. Barr et al., 2013), except for the correlation between the intercept and the slope for empirical salience, which was the predictor of interest for the present study. Mathematical details of the models are provided in Appendix C.

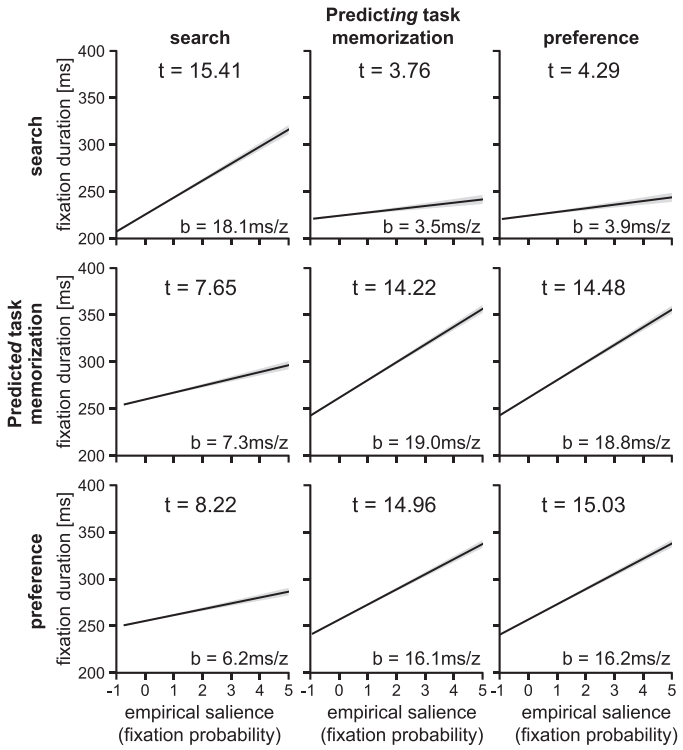

Figure 5. Experiment 1: Prediction of fixation durations by fixation probability after statistically controlling for low-level feature effects. Partial effect of fixation probability on fixation duration, in models that include image features as additional fixed effects. Notation as in Figure 4.

Assessing statistical significance

Since the degrees of freedom are ill-defined in LMMs (Baayen, Davidson, & Bates, 2008), there is no closed-form estimate that yields a p value. If the number of observations (here, fixations) is large compared to the number of predictors, for all practical purposes the t distribution is identical to the normal distribution (Baayen et al., 2008). Hence, an absolute t value greater than 1.96 corresponds to a two-sided significance criterion at a 5% alpha level.

Implementation

All image processing was done using MATLAB version R2007b; all statistical analysis was conducted using the R system for statistical computing (version 3.1; R Core Team, 2014). Linear mixed-effects models were computed using the lmer program of the lme4 R package (version 1.1.7; Bates, Maechler, Bolker, & Walker, 2014) with the bobyqa optimizer and a maximum of 10^7 function evaluations. Model parameters were estimated through restricted maximum-likelihood estimation. For computation of partial LMM effects, the remef function (Hohenstein & Kliegl, 2014) was used.

Results

Experiment 1: Fixation probability predicts fixation duration across tasks

In Experiment 1, each scene was viewed by 72 observes in one of three viewing tasks (24 observers per task). Each task could serve as predicting condition (the condition that fixation probability is estimated from) and as predicted condition. The combination of three predicting conditions with three predicted conditions resulted in a total of nine individual models. For all nine models, fixation probability significantly predicted fixation duration, such that locations that were frequently selected for fixation also received longer fixations (Figure 4 for all t values). For each model, the regression coefficient b quantifies the increase of fixation duration (in milliseconds) per standard-deviation (z-unit) increase in empirical salience; it reached a maximum of 19.6 ms/z (Figure 4 for all values of b).

For six of the models, the predicted condition was different from the predicting condition (off-diagonal models in Figure 4). In these cases, the predicting and predicted sets of observers were entirely independent for any given item, by virtue of the experiment's design. In general, effects were smaller when search was either the predicting or predicted task than for predictions not involving search. For the remaining three models (on-diagonal in Figure 4), predicted and predicting condition were identical. For these models, the empirical salience map that predicted fixation durations for a given individual and image also included this individual. To address this issue, we recomputed the three on-diagonal models with leave-one-out empirical salience maps: For each observation in the model (i.e., each fixation), the computation of empirical salience excluded the corresponding observer. This leave-one-out empirical salience still significantly predicted fixation duration (search: $b = 16.1$ ms/z, $t = 13.56$; memorization: $b = 17.7$ ms/z, $t = 13.56$; preference: $b = 14.9$ ms/z, $t = 14.11$). Hence, prediction of duration by probability in the same task was not dominated by observer idiosyncrasies. In sum, the results suggest that fixation probability in an image predicts fixation duration, and it does so across tasks and in independent observers.

Fixation probability predicts duration above and beyond image features or central bias

The result that fixation probability predicts fixation duration could potentially depend on low-level image features, if there is a low-level feature that relates to fixation probability and fixation duration alike. A similar confound could arise if fixation duration and

fixation probability were to exhibit a similar dependence on the spatial location within the scene. To address this issue, we included several features in the model that are known to affect fixation probability or duration (Figure 2). We again specified nine LMMs, which now included the three image features and the eccentricity variable as additional fixed effects along with the empirical-salience predictor; details on the random-effects structure are provided in Materials and methods. Including the additional predictors did not affect the overall result pattern. For all nine models, fixation probability—as measured by empirical salience—remained a significant predictor of fixation duration (Figure 5). As with the simple models, among the six off-diagonal models prediction appeared to be best for those predictions not involving search as either the predicted or predicting condition. Again, we recomputed the three on-diagonal models (i.e., identical predicted and predicting condition) with leave-one-out empirical salience maps and found that the significant prediction prevailed (search: $b = 14.7$ ms/z, $t = 12.95$; memorization: $b = 17.8$ ms/z, $t = 13.54$; preference: $b = 15.2$ ms/z, $t = 14.24$). The results suggest that neither any of the tested features nor central bias can explain the observed relation between fixation duration and fixation probability.

Factors guiding fixations may change over viewing time. It is a well-established finding that fixation durations increase during initial viewing periods and stabilize during later viewing (for a review, see Nuthmann, 2016). Moreover, effects of low-level features on fixation probability appear to decline over time (Parkhurst, Law, & Niebur, 2002); some of this decline can be explained by spatial biases (Tatler, Baddeley, & Gilchrist, 2005), and both factors become less important during prolonged viewing (Wang et al., 2015). The question therefore arises whether the relation of fixation probability and fixation duration takes substantial time to develop. To test this, we ran additional analyses that considered only the first n seconds of the experiment. For most conditions, the effect in question quickly reaches the pattern observed for the full viewing time (Figure 6). The prediction of search by itself takes about 5 s to reach its asymptotic level. This time (5 s) is chosen as the presentation duration for Experiment 2.

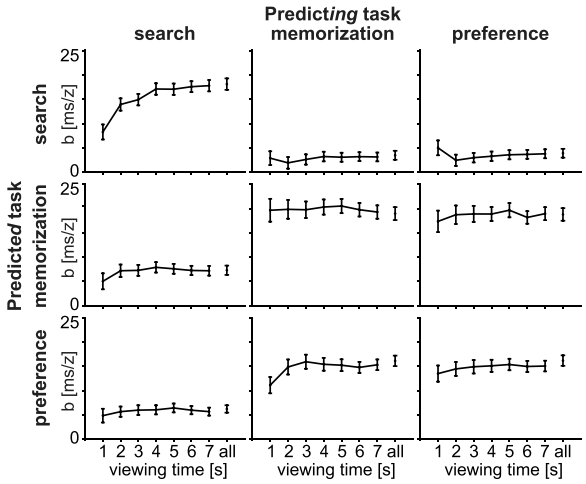

Figure 6. Time-course analysis of the relation between fixation duration and fixation probability in Experiment 1. Standardized fixed-effects regression coefficients are plotted against integrated viewing time. Error bars depict the standard errors of the estimates. All regression slopes are significantly larger than 0 ($t > 1.96$), except the prediction of the search task by memorization after 2 s. In each panel, the rightmost data point ("all") corresponds to the data in Figure 5.

Experiment 2: The relation of fixation probability and duration is insensitive to the experimental manipulation of low-level features

Mixed-effects modeling of the data from Experiment 1 suggests that the relation between fixation probability and fixation duration is insensitive to low-level stimulus features. In Experiment 2, we followed the LMM results up with experimental manipulations. Does the prediction persist if image features are modified experimentally, and if participants are asked to freely view the images?

Stimuli from Condition 1 in Experiment 2 matched a subset from Experiment 1, except with color removed. For these grayscale images, fixation probability still predicts fixation duration ($b = 16.2$ ms/z; $t = 8.95$; Figure 7 upper left panel). In Conditions 2 through 5, in addition to removing color, we reduced or equalized other low-level features (luminance, contrast; first five rows and columns of Figure 7). Under these experimental modifications, the prediction also persists: Both within a condition (diagonal in Figure 7) and across conditions (off-diagonal panels in Figure 7), the regression coefficients range between 11.3 and 18.7 ms per z-normalized empirical salience value (all ts > 6.6), and this range is comparable to the non-search conditions of Experiment 1 (Figure 5). Hence, the experimental manipulations of Experiment 2 confirm the statistical analysis of Experiment 1: The prediction of fixation duration by fixation probability persists when features are accounted for. In contrast, when higher level information is destroyed (phase-noise Condition 6), fixation duration is not predicted by fixation probability (Figure 7, bottom row and right column, all $|t| < 1.35$). Interestingly, fixation probability in the phase-noise condition predicts fixation duration in the phase-noise condition, suggesting that observers interpret consistent structure in the noise. As

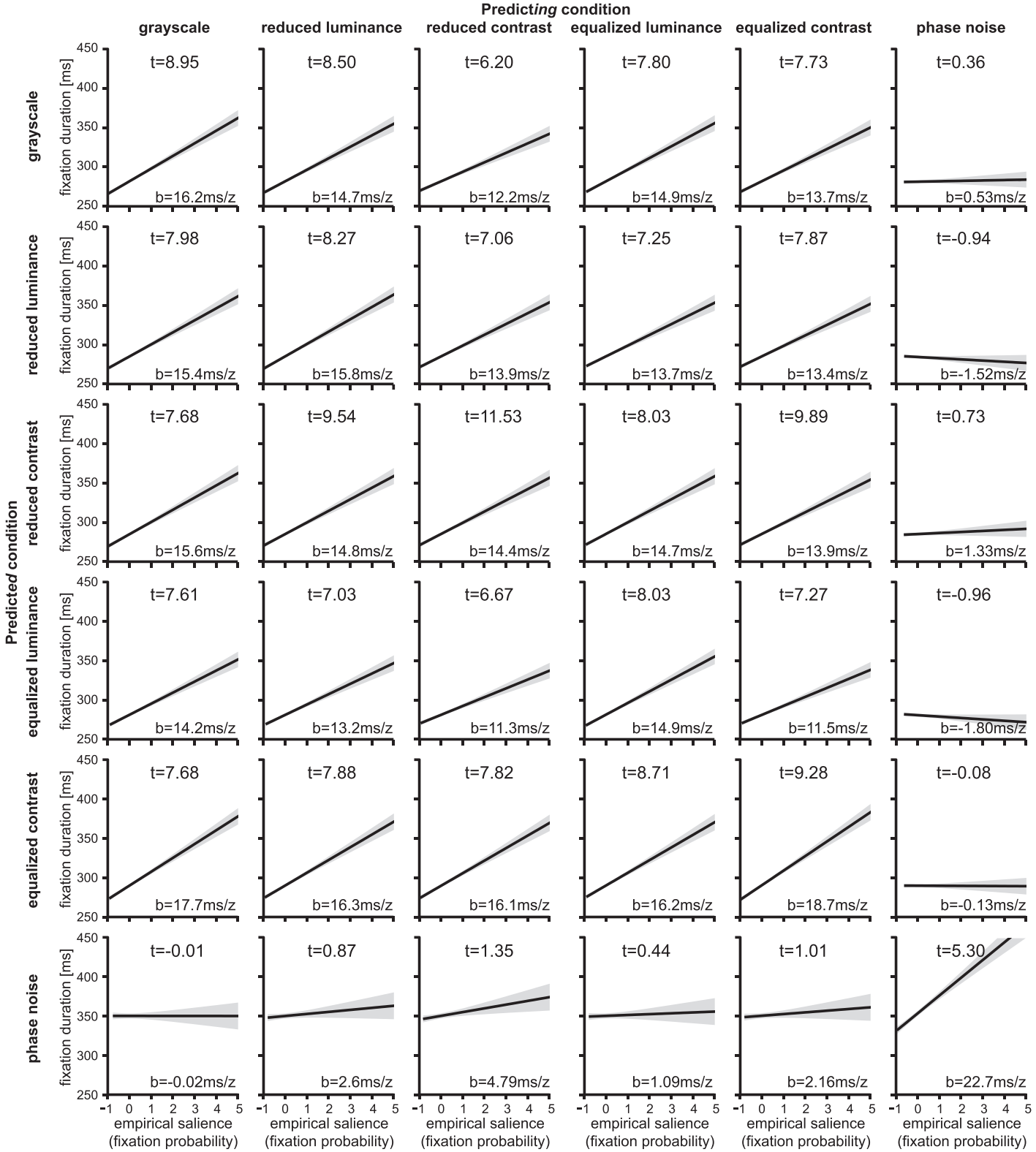

Figure 7. Experiment 2. Prediction of fixation duration by empirical salience in Experiment 2. The predicting condition from which the empirical map is computed is given by the column, the predicted condition by the row. Notation as in Figure 4.

for the analysis of Experiment 1, we repeat the analysis but now exclude the participant whose durations are considered from the computation of the corresponding empirical map (leave-one-out prediction). The results are consistent with the main analysis: For all 25 pairs of predicting and predicted condition that exclude phase noise, there is a significant prediction (all ts > 6.35, bs between 10.7 and 17.3 ms/z). The same is true for the prediction of phase noise by itself ($b = 19.8$ ms/z; $t = 4.84$; Figure 7), while for the combination of phase noise and other conditions, no such prediction is observed (all $|t| < 1.1$). As every image is presented to each observer once in each condition, each observer sees each image six times (albeit in different versions). To discount the possibility that the results are affected by those item repetitions, we repeated the analysis using data from the first block only, which still provides us with a fully balanced between-subjects design per image and condition. We find the same qualitative result: For the 25 models excluding the phase-noise condition as predicted and predicting condition, fixation probability predicts fixation duration (all ts > 2.6); for the 10 models including phase noise as either predicted or predicting condition, no such effect is found (all $|t| < 1.3$). The "self-prediction" of phase noise is again significant ($t = 5.0$).

In sum, these data show that the prediction of fixation duration by fixation probability is not explained by the low-level features (luminance, contrast) that change between the first five conditions, but by (higher level) properties that are changed only in the phase-noise condition.

Fixation probability in Experiment 1 predicts fixation duration in Experiment 2

Finally, we asked how well fixation probability from the corpus data predicts fixation durations in Experiment 2. Notably, in all conditions of Experiment 2, images were deprived of color, while in Experiment 1 color stimuli were used. Besides robustness across labs, setups, and cultural background, successful prediction of Experiment 2 by Experiment 1 would therefore demonstrate that the presence of color is not a necessary condition for the relation between fixation duration and fixation probability.

We used the empirical maps of Experiment 1 to predict fixation durations in Experiment 2. For each model, the same LMM model structure was applied as the one used for the data presented in Figures 5 through 7. The empirical maps from any task in Experiment 1 significantly predicted fixation duration in all five conditions of Experiment 2, in which only low-level features were modified (Figure 8, top five rows). The standardized regression coefficients indicate that, numerically, the fixation probability obtained from memorization and preference predict fixation duration in free viewing somewhat better than fixation probability obtained from search. When higher level scene structure was removed, the prediction vanished (Figure 8, bottom). This confirms across experiments that the relation between fixation probability and fixation duration is insensitive to the removal or modification of low-level features, including color besides luminance and contrast, as long as higher level scene structure remains intact.

Discussion

The present study demonstrated a systematic relationship between fixation probability and fixation duration in real-world scene perception and search: The locations that observers were more likely to select for fixation were also the locations that received longer fixations. We provided two lines of evidence that this key result was not confounded by low- and mid-level luminance features or by an eccentricity bias. First, we used a statistical control approach that allowed for assessing the *independent* contribution of fixation probability to fixation duration; second, we showed experimentally that fixation probabilities obtained in the original scenes continued to predict fixation durations in scenes deprived of low-level features.

In the present study we used four distinct tasks: search, memorization, and preference judgment in Experiment 1, and free viewing in Experiment 2. Tasks varied in the degree of top-down control they typically evoke (cf. Nuthmann & Matthias, 2014), ranging from minimal top-down influences in free viewing to strong top-down control in visual search. Viewing task can have a profound influence on fixation probability (Buswell, 1935; Yarbus, 1967). In particular, search for a target reduces or abolishes the effect of bottom-up signals (Henderson, Brockmole, Castelhano, & Mack, 2007; Einhäuser, Rutishauser, & Koch, 2008). Perhaps not surprisingly, predictions across tasks were worse if search was involved as either predicting or predicted condition (Figures 4, 5, and 8). Nonetheless, fixation durations in search were still significantly predicted by fixation probabilities in the other tasks, and in turn, empirical salience from search significantly predicted fixation durations in the other tasks. These results indicate that part of the relation between fixation probability and fixation duration is insensitive to task, at least for the tasks tested.

No analysis can exclude *all* possible features that could potentially contribute to the relation of fixation duration and fixation probability, which is why we restricted ourselves to the most commonly used

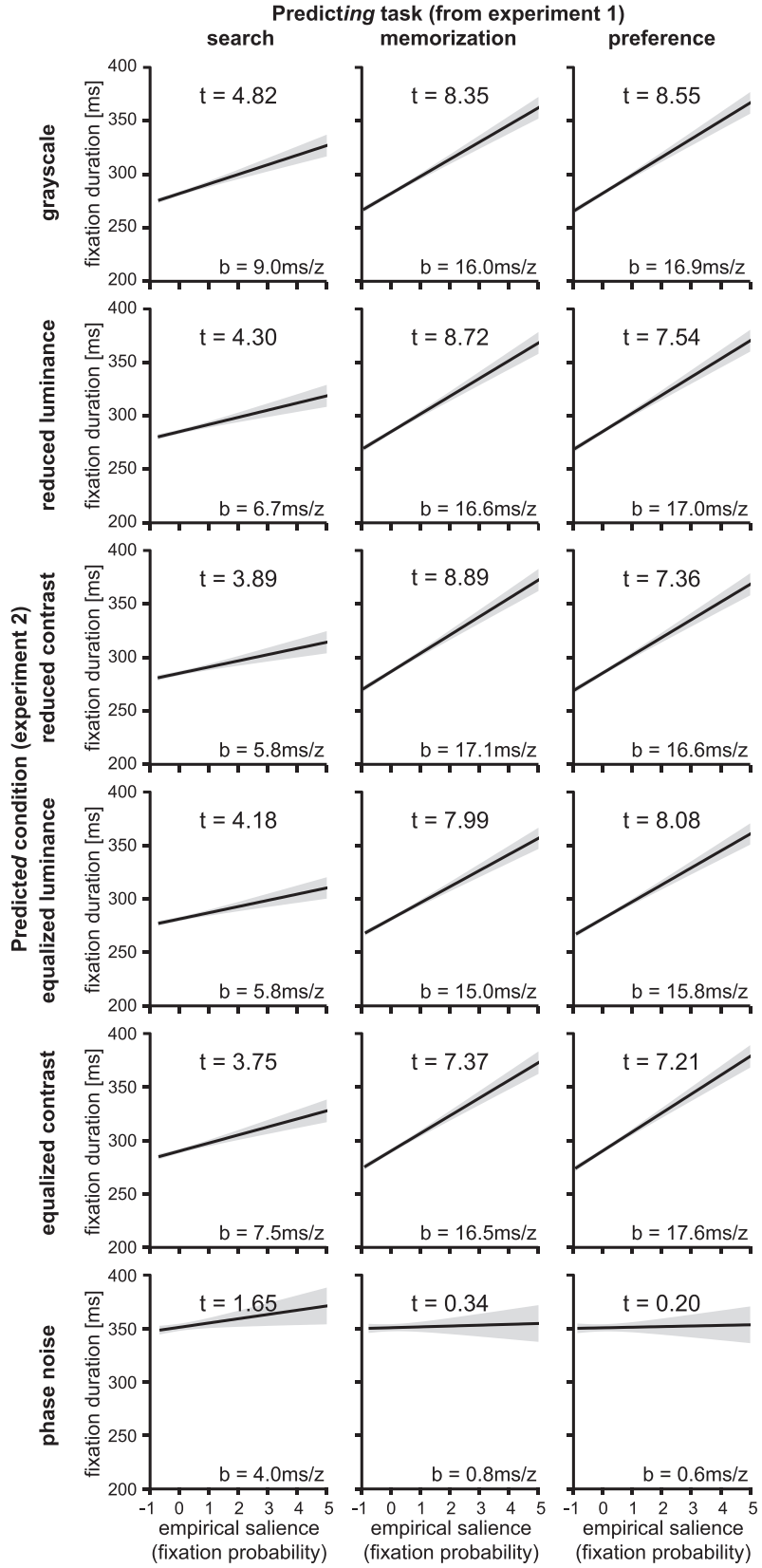

Figure 8. Fixation probabilities obtained in Experiment 1 predict fixation durations in Experiment 2: Partial effect of fixation probability on fixation duration. Empirical maps from search task of Experiment 1 (left column), memorization task (middle column), and preference-judgment task (right column) predict fixation durations for the different stimulus conditions of Experiment 2 (rows).

achromatic low- and mid-level features. We deliberately refrained from including chromatic features, as the corpus stimuli were not controlled for veridical color representation. The role of color in guiding attention during scene viewing is indeed debated and most likely task dependent. In free viewing, Tatler et al. (2005) found only a small effect of chromaticity as compared to contrast and edge density. Similarly, during scene memorization, removing color from the scene images had little effect on fixation patterns (Harding & Bloj, 2010). Frey et al. (2011) manipulated color information in scenes during free viewing in normal observers and those with color deficiency. They found that contrasts along the red–green axis relate to fixation probability, but their influence does not seem to be causal. In contrast, for a search task, chromatic information may supersede achromatic information (Amano & Foster, 2014). Although we did not include chromatic features in our statistical models, Experiment 2 allows us to discount the possibility that the relation between fixation duration and fixation probability is exclusively or primarily driven by color. Fixation probability from the corpus study (Experiment 1), in which the images were presented in color, predicts fixation probability in Experiment 2, in which all stimuli were presented in grayscale—that is, without chromatic information. Nonetheless, the systematic experimental manipulation of chromatic features with well-defined chromatic stimuli remains an important issue for further research.

By design, our analysis only addresses fixated areas. Although—by normalization of the empirical maps—fixation probability is defined relative to the whole scene, fixation duration can only be assessed for locations that are actually fixated. It can thus not be ruled out that the non-visited regions of a scene would act contrary to prediction (i.e., produce longer fixations) if fixations there would be enforced. Although this might be an interesting option for further experimental investigation, it is clearly beyond the scope of the present study.

In this article, we combine corpus-based analyses with experimental manipulations: The *correlational* results obtained from the corpus data (Experiment 1) guided the *experimental* design of Experiment 2. Although such a combined strategy is rather common in the research on eye-movement control in reading (Angele et al., 2015; Kliegl, 2007; Kliegl, Nuthmann, & Engbert, 2006; Rayner, Pollatsek, Drieghe, Slattery, & Reichle, 2007), it has rarely been followed in natural scene viewing.

In contrast to Buswell's hypothesis, the separation of the *when* from the *where* of fixation selection has been the dominant view in natural scene viewing over the last decades. It has influenced the development of theoretical and computational models as well as the design of behavioral and physiological experiments. The when/where distinction has led to the development of increasingly sophisticated salience models (Borji & Itti, 2013; Borji et al., 2013a) which ignore fixation durations and has fostered the development of a theoretical model of fixation duration (Nuthmann et al., 2010) which does not currently implement a target-selection mechanism. Similarly, experimental studies addressed *either* the role of local image features in the selection of fixated locations (e.g., Baddeley & Tatler, 2006, Borji & Itti, 2013; Reinagel & Zador, 1999) or the effect of image-wide feature modifications on fixation duration (Henderson et al., 2013; Ho-Phuoc et al., 2012; Kaspar & König, 2011; Walshe & Nuthmann, 2014, 2015). Recent work has quantified the independent effects of *local* image features on fixation duration (Nuthmann, 2016), akin to their effects on fixation probability (Nuthmann & Einhäuser, 2015): All image features considered here (luminance, contrast, edge density) were shown to have an independent contribution to fixation duration in mixed models that controlled for spatial and oculomotor constraints. All of these studies considered *either* fixation duration *or* fixation probability. Besides conceptual (Findlay & Walker, 1999) and neurobiological (van Gisbergen et al., 1981) reasons, this choice may also result from practical considerations: Especially in considering several factors and features in parallel, the when/where division helps to keep experiments, regression analyses, and computational modeling tractable. Here we used empirical salience to estimate fixation probability. Linear mixed-effects modeling showed that fixation probability had a significant effect on fixation duration, even after controlling for local image features and central bias. We then went beyond a pure statistical control approach and controlled low-level features experimentally. The results from Experiment 2 confirmed that the when/where relation was insensitive to image features and generalized over tasks, setups, and labs. Eighty years after Buswell's original proposal, we now have the analytical and experimental tools available to establish a systematic relationship between fixation probability and duration. At least on the level of behavior, these findings challenge the prevalent notion of a separation between when and where decisions of attentional selection.

As a parsimonious explanation for our results, we suggest that it is some form of "interestingness" or "relevance" that makes a location more likely to be fixated *and* fixated longer. Indeed, several studies found that regions of a scene that had been labeled as more interesting, more relevant, or more informative were also more likely to be fixated (Antes, 1974; Mackworth & Morandi, 1967; Masciocchi et al., 2009; Onat et al., 2014). In one of the very few studies that related such subjective measures of relevance to fixation duration,

Onat et al. found a positive correlation between "interestingness" and fixation durations. They also report a correlation between interestingness and fixation probability, but they did not assess the relationship between fixation probability and duration. The results from Experiment 2 lend support to the role of interestingness in explaining the relation between the *where* and the *when*. If low-level features were degraded in a scene, the relation persisted; if, however, higher level structure was removed (phase-noise condition), the relation vanished. Phase randomization destroys not only objects but also structure that may gain object-like properties (gestalt) without necessitating a semantic interpretation. Using artificial stimuli, it has been demonstrated that such "perceptual objects" can guide attention (Yeshurun, Kimchi, Sha'shoua, & Carmel, 2009). In the light of these findings, interestingness in natural scenes does not necessarily require semantic meaning. Instead, it may refer to any scene property to which distinct observers consistently attribute some form of perceptual relevance. Such consistent attribution of relevance may also underlie the prediction of fixation duration by probability in the phase-noise condition: Here, large-scale noise structure may consistently be interpreted as some form of content. Such persistence of second-order structure across spatial scales is a property that distinguishes $1/f$ noise from other types of noise (e.g., white noise). It will therefore be an interesting question for further research to investigate for which types of noise a relation between fixation probability and fixation duration can be observed. The interestingness interpretation suggested by the present data is in line with Buswell's original hypothesis on the "centers of interest" receiving higher probability and higher duration. Building upon our results, further research may directly manipulate interestingness to test how it affects fixation duration, fixation probability, and their relation. Moreover, our results inform and constrain computational models of fixation guidance. First, a full model needs to explain both fixation probability and fixation duration. Second, it needs to achieve a level of "scene understanding" that captures those aspects of higher level scene content that control gaze in space *and* time.

Keywords: visual attention, scene viewing, natural scene, eye movements, gaze

Acknowledgments

The authors thank D. Walper for support in data collection for Experiment 2 and the German Research Foundation (DFG; SFB/TRR135) for financial support.

Commercial relationships: none.

Corresponding author: Wolfgang Einhäuser.
Email: wolfgang.einhaeuser-treyer@physik.tu-chemnitz.de.
Address: Technische Universität Chemnitz, Institute of Physics, Chemnitz, Germany.

Footnote

[1] Data from this corpus have previously been used to study attentional selection within objects (Nuthmann & Henderson, 2010; Pajak & Nuthmann, 2013) and specific viewing biases during scene perception (Luke et al., 2014; Nuthmann & Matthias, 2014). In contrast, here the data are used—in addition to new experimental data—to assess the relationship between fixation probability and fixation duration, which has not been addressed before.

References

Amano, K., & Foster, D. H. (2014). Influence of local scene color on fixation position in visual search. *Journal of the Optical Society of America A: Optics, Image Science, and Vision*, *31*(4), A254–A262.

Angele, B., Schotter, E. R., Slattery, T. J., Tenenbaum, T. L., Bicknell, K., & Rayner, K. (2015). Do successor effects in reading reflect lexical parafoveal processing? Evidence from corpus-based and experimental eye movement data. *Journal of Memory and Language*, *79*, 76–96.

Antes, J. R. (1974). The time course of picture viewing. *Journal of Experimental Psychology*, *103*, 62–70.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Baddeley, R. J., & Tatler, B. W. (2006). High frequency edges (but not contrast) predict where we fixate: A Bayesian system identification analysis. *Vision Research*, *46*, 2824–2833.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, D. M., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4 (R package version 1.1-7). Retrieved from http://CRAN.Rproject.org/package=lme4

Bex, P. J., & Makous, W. (2002). Spatial frequency, phase, and the contrast of natural images. *Journal of*

the Optical Society of America A: Optics, Image Science, and Vision, *19*(6), 1096–1106.

Borji, L., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 185–207.

Borji, A., Sihite, D. N., & Itti, L. (2013a). Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing*, *22*(1), 55–69.

Borji, A., Sihite, D. N., & Itti, L. (2013b). What stands out in a scene? A study of human explicit saliency judgment. *Vision Research*, *91*, 62–77.

Bruce, N. D., & Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, *9*(3):5, 1–24, doi: 10.1167/9.3.5.

Buswell, G. T. (1935). *How people look at pictures. A study of the psychology of perception in art*. Chicago: University of Chicago Press.

Clarke, A. D., & Tatler, B. W. (2014). Deriving an appropriate baseline for describing fixation behaviour. *Vision Research*, *102*, 41–51.

Dakin, S. C., Hess, R. F., Ledgeway, T., & Achtman, R. L. (2002). What causes non-monotonic tuning of fMRI response to noisy images? *Current Biology*, *12*, R476–R477.

Einhäuser, W., & König, P. (2003). Does luminance-contrast contribute to a saliency map for overt visual attention? *European Journal of Neuroscience*, *17*, 1089–1097.

Einhäuser, W., Rutishauser, U., Frady, E. P., Nadler, S., König, P., & Koch, C. (2006). The relation of phase noise and luminance contrast to overt attention in complex visual stimuli. *Journal of Vision*, *6*(11):1, 1148–1158, doi:10.1167/6.11.1. [PubMed] [Article]

Einhäuser, W., Rutishauser, U., & Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision*, *8*(2):2, 1–19, doi:10.1167/8.2.2. [PubMed] [Article]

Elazary, L., & Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision*, *8*(3):3, 1–15, doi: 10.1167/8.3.3. [PubMed] [Article]

Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review*, *112*(4), 777–813.

Erdem, E., & Erdem, A. (2013). Visual saliency estimation by nonlinearly integrating features using region covariances. *Journal of Vision*, *13*(4):11, 1–20, doi:10.1167/13.4.11. [PubMed] [Article]

Findlay, J. M., & Walker, R. (1999). A model of saccade generation based on parallel processing and competitive inhibition. *Behavioral and Brain Sciences*, *22*(4), 661–674.

Frey, H. P., Wirz, K., Willenbockel, V., Betz, T., Schreiber, C., Troscianko, T., & König, P. (2011). Beyond correlation: Do color features influence attention in rainforest? *Frontiers in Human Neuroscience*, *5*, 36.

Garcia-Diaz, A., Leborán, V., Fdez-Vidal, X. R., & Pardo, X. M. (2012). On the relationship between optical variability, visual saliency, and eye fixations: A computational approach. *Journal of Vision*, *12*(6): 17, 1–22, doi:10.1167/12.6.17. [PubMed] [Article]

Harding, G., & Bloj, M. (2010). Real and predicted influence of image manipulations on eye movements during scene recognition. *Journal of Vision*, *10*(2):8, 1–17, doi:10.1167/10.2.8. [PubMed] [Article]

Harel, J., Koch, C., & Perona, P. (2007). Graph-based visual saliency. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems 19 (NIPS 2006)* (pp. 545–552). Cambridge, MA: MIT Press.

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. (2007). Visual saliency does not account for eye-movements during visual search in real-world scenes. In R. Van Gompel, M. Fischer, W. Murray, & R. Hills (Eds.), *Eye movement research: Insights into mind and brain* (pp. 437–562). Oxford, UK: Elsevier.

Henderson, J. M., & Choi, W. (2015). Neural correlates of fixation duration during real-world scene viewing: Evidence from fixation-related (FIRE) fMRI. *Journal of Cognitive Neuroscience*, *27*(6), 1137–1145.

Henderson, J. M., Nuthmann, A., & Luke, S. G. (2013). Eye movement control during scene viewing: Immediate effects of scene luminance on fixation durations. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(2), 318–322.

Ho-Phuoc, T., Guyader, N., Landragin, F., & Guérin-Dugué, A. (2012). When viewing natural scenes, do abnormal colors impact on spatial or temporal parameters of eye movements? *Journal of Vision*, *12*(2):4, 1–13, doi:10.1167/12.2.4. [PubMed] [Article]

Hohenstein, S., & Kliegl, R. (2014). Semantic preview benefit during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(1), 166–190.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*, 1489–1506.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual-attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*, 1254–1259.

Kaspar, K., & König, P. (2011). Overt attention and

context factors: The impact of repeated presentations, image type, and individual motivation. *PLoS ONE*, *6*(7), e21719.

Kayser, C., Nielsen, K. J., & Logothetis, N. K. (2006). Fixations in natural scenes: Interaction of image structure and image content. *Vision Research*, *46*, 2535–2545.

Kliegl, R. (2007). Toward a perceptual-span theory of distributed processing in reading: A reply to Rayner, Pollatsek, Drieghe, Slattery, and Reichle (2007). *Journal of Experimental Psychology: General*, *136*(3), 530–537.

Kliegl, R., Nuthmann, A., & Engbert, R. (2006). Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General*, *135*(1), 12–35.

Koehler, K., Guo, F., Zhang, S., & Eckstein, M. P. (2014). What do saliency models predict? *Journal of Vision*, *14*(3):14, 1–27, doi:10.1167/14.3.14. [PubMed] [Article]

Kümmerer, M., Wallis, T. S. A., & Bethge, M. (2015). Information-theoretic model comparison unifies saliency metrics. *Proceedings of the National Academy of Sciences, USA*, *112*(52), 16054–16059, doi:10.1073/pnas.1510393112.

Lin, R. J., & Lin, W. S. (2014). A computational visual saliency model based on statistics and machine learning. *Journal of Vision*, *14*(9):1, 1–18, doi:10.1167/14.9.1. [PubMed] [Article]

Luke, S. G., Smith, T. J., Schmidt, J., & Henderson, J. M. (2014). Dissociating temporal inhibition of return and saccadic momentum across multiple eye-movement tasks. *Journal of Vision*, *14*(14):9, 1–12, doi:10.1167/14.14.9. [PubMed] [Article]

Mackworth, N. H., & Morandi, A. J. (1967). The gaze selects informative details within pictures. *Perception and Psychophysics*, *2*, 547–552.

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1996). The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial Vision*, *10*, 165–188.

Masciocchi, C. M., Mihalas, S., Parkhurst, D., & Niebur, E. (2009). Everyone knows what is interesting: Salient locations which should be fixated. *Journal of Vision*, *9*(11):25, 1–22, doi:10.1167/9.11.25. [PubMed] [Article]

Moulden, B., Kingdom, F., & Gatley, L. F. (1990). The standard deviation of luminance as a metric for contrast in random-dot images. *Perception*, *19*(1), 79–101.

Nuthmann, A. (2016). Fixation durations in scene viewing: Modeling the effects of local image features, oculomotor parameters, and task. *Psychonomic Bulletin & Review*, 1–23, doi:10.3758/s13423-016-1124-4.

Nuthmann, A., & Einhäuser, W. (2015). A new approach to modeling the influence of image features on fixation selection in scenes. *Annals of the New York Academy of Sciences*, *1339*, 82–96.

Nuthmann, A., & Henderson, J. M. (2010). Object-based attentional selection in scene viewing. *Journal of Vision*, *10*(8):20, 1–19, doi:10.1167/10.8.20. [PubMed] [Article]

Nuthmann, A., & Henderson, J. M. (2012). Using CRISP to model global characteristics of fixation durations in scene viewing and reading with a common mechanism. *Visual Cognition*, *20*(4–5), 457–494.

Nuthmann, A., & Malcolm, G. L. (2016). Eye-guidance during real-world scene search: The role color plays in central and peripheral vision. *Journal of Vision*, *16*(2):3, 1–16, doi:10.1167/16.2.3. [PubMed] [Article]

Nuthmann, A., & Matthias, E. (2014). Time course of pseudoneglect in scene viewing. *Cortex*, *52*, 113–119.

Nuthmann, A., Smith, T. J., Engbert, R., & Henderson, J. M. (2010). CRISP: A computational model of fixation durations in scene viewing. *Psychological Review*, *117*(2), 382–405.

Onat, S., Açık, A., Schumann, F., & König, P. (2014). The contributions of image content and behavioral relevancy to overt attention. *PLoS ONE*, *9*(4), e93254.

Pajak, M., & Nuthmann, A. (2013). Object-based saccadic selection during scene perception: evidence from viewing position effects. *Journal of Vision*, *13*(5):2, 1–21, doi:10.1167/13.5.2. [PubMed] [Article]

Parkhurst, D. J., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, *42*(1), 107–123.

Pomplun, M., Ritter, H., & Velichkovsky, B. (1996). Disambiguating complex visual information: Towards communication of personal views of a scene. *Perception*, *25*(8), 931–948.

Rainer, G., Augath, M., Trinath, T., & Logothetis, N. K. (2001). Nonmonotonic noise tuning of BOLD fMRI signal to natural images in the visual cortex of the anesthetized monkey. *Current Biology*, *11*, 846–854.

Rainer, G., Lee, H., & Logothetis, N. K. (2004). The effect of learning on the function of monkey extrastriate visual cortex. *PLoS Biology*, *2*, E44.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422.

Rayner, K., Pollatsek, A., Drieghe, D., Slattery, T. J., & Reichle, E. D. (2007). Tracking the mind during reading via eye movements: Comments on Kliegl, Nuthmann, and Engbert (2006). *Journal of Experimental Psychology: General*, *136*(3), 520–529.

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The E-Z reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences*, *26*(4), 445–476.

Reinagel, P., & Zador, A. M. (1999). Natural scene statistics at the centre of gaze. *Network: Computation in Neural Systems*, *10*, 341–350.

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Schielzeth, H. (2010). Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution*, *1*(2), 103–113.

Stoll, J., Thrun, M., Nuthmann, A., & Einhäuser, W. (2015). Overt attention in natural scenes: Objects dominate features. *Vision Research*, *107*, 36–48.

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, *7*(14):4, 1–17, doi: 10.1167/7.14.4. [PubMed] [Article]

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research*, *45*, 643–659.

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision reinterpreting salience. *Journal of Vision*, *11*(5):5, 1–23, doi:10.1167/11.5.5. [PubMed] [Article]

van Gisbergen, J., Gielen, S., Cox, H., Bruijns, J., & Kleine Schaars, H. (1981). Relation between metrics of saccades and stimulus trajectory in visual target tracking: Implications for models of the saccadic system. In A. F. Fuchs & W. Becker (Eds.), *Progress in oculomotor research* (pp. 19–27). Amsterdam: North Holland, Elsevier.

Vincent, B. T., Baddeley, R., Correani, A., Troscianko, T., & Leonard, U. (2009). Do we look at lights? Using mixture modelling to distinguish between low- and high-level factors in natural image viewing. *Visual Cognition*, *17*, 856–879.

Walshe, R. C., & Nuthmann, A. (2014). Asymmetrical control of fixation durations in scene viewing. *Vision Research*, *100*, 38–46.

Walshe, R. C., & Nuthmann, A. (2015). Mechanisms of saccadic decision making while encoding naturalistic scenes. *Journal of Vision*, *15*(5):21, 1–19, doi:10.1167/15.5.21. [PubMed] [Article]

Wang, S., Jiang, M., Duchesne, X. M., Laugeson, E. A., Kennedy, D. P., Adolphs, R., & Zhao, Q. (2015). Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking. *Neuron*, *88*(3), 604–616.

Wichmann, F. A., Braun, D. I., & Gegenfurtner, K. R. (2006). Phase noise and the classification of natural images. *Vision Research*, *46*, 1520–1529.

Xu, J., Jiang, M., Wang, S., Kankanhalli, M. S., & Zhao, Q. (2014). Predicting human gaze beyond pixels. *Journal of Vision*, *14*(1):28, 1–20, doi:10.1167/14.1.28. [PubMed] [Article]

Yarbus, A. L. (1967). *Eye movements and vision*. New York: Plenum.

Yeshurun, Y., Kimchi, R., Sha'shoua, G., & Carmel, T. (2009). Perceptual objects capture attention. *Vision Research*, *49*, 1329–1335.

Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, *8*(7):32, 1–20, doi:10.1167/8.7.32. [PubMed] [Article]

Appendix A: Mathematical definition of low-level features

Three local image features were computed for each stimulus as described in the following: local luminance, local luminance contrast with global normalization, and edge density. For the data of Experiment 2, the actually displayed luminance, ranging from 0.1 to 66.0 cd/m^2, was used for computing the features. For the data of Experiment 1, no information about the screen luminance was available, and the intensity (pixel value after transforming the image to grayscale using MATLAB's rgb2gray function, ranging from 0 to 255) was used in lieu of luminance. For simplicity of notation, we denote the respective values at each point for either definition as $I(x, y)$. For all features, a Gaussian kernel describing the local region was defined as

$$G(x,y) = \frac{\exp\left[-\frac{x^2+y^2}{2\sigma^2}\right]}{2\pi\sigma^2}.$$

In line with the empirical maps, $\sigma = 16$ pixels (corresponding to 0.5° of visual angle) was used. For computational purposes, the Gaussian was restricted to the patch 81×81 pixels wide and normalized to unit integral within this patch.

To minimize edge effects in computing features close to the image boundary, the image was extended to each side with a mirrored version (mirrored at the respective image border), and the result of convolutions was cropped to the original image size. Only

fixations that fell inside the image boundaries were used for analysis.

Local luminance

The local luminance was defined as the weighted mean luminance around a given location:

$$L = G*I,$$

where * denotes the convolution.

Luminance contrast

For luminance contrast, we first defined the Gaussian-weighted variance as

$$V = (G*I^2) - (G*I)^2,$$

where $()^2$ denotes point-wise squaring of each pixel.

The contrast $C(x, y)$ was then obtained by normalizing the square-root of this variance by the image mean $<L(x, y)>$:

$$C(x,y) = \frac{\sqrt{V(x,y)}}{L(x,y)}.$$

Edge density

Using a Sobel filter and image-specific thresholds as given by the default settings of MATLAB's edge function, we computed an edge image X from the original image. To obtain the weighted edge density in a local environment, the edge image was convolved with the Gaussian kernel, averaging the number of edges in the local region:

$$E = G*X.$$

Relation to other measures of luminance, luminance contrast, and edge density

We used the Gaussian weighting for consistency between the definition of features and the empirical maps. It should be noted that replacing the kernel G in the previous definitions with the uniform kernel

$$U(x,y) = \begin{cases} \frac{1}{w^2} & -\frac{w}{2} \le x,y \le \frac{w}{2} \\ 0 & \text{otherwise} \end{cases}$$

with $w = 32$ pixels yields the widely used definitions of luminance, edge density, and luminance contrast. With the uniform kernel, one obtains the average luminance in a 1° square, the edge density in a 1° square (Baddeley & Tatler, 2006), and, since the relation

$$\begin{aligned} \mathrm{var}(x) &= \frac{1}{N-1}\sum (x_i - \bar{x})^2 \\ &= \frac{N}{N-1}\left(\frac{\sum x_i^2}{N} - \bar{x}^2\right) \approx \frac{\sum x_i^2}{N} - \bar{x}^2 \end{aligned}$$

allows the (unweighted) variance to be written as

$$V_u = (U*I^2) - (U*I)^2,$$

the root-mean-square contrast in a 1° square (Reinagel & Zador, 1999).

We note that some salience models use definitions of contrast that are closer to the center–surround organization in the earliest stages of vision (e.g., Itti & Koch, 2000). We chose to use a variant of root-mean-square contrast, since it is consistent with measures of detectability in natural scenes (Bex & Makous, 2002) and with filter properties of early vision (Moulden, Kingdom, & Gatley, 1990) and because it has been used in previous experimental studies on fixation selection in scenes (Einhäuser & König, 2003; Mannan et al., 1996; Reinagel & Zador, 1999).

Appendix B: Stimulus modifications in Experiment 2

In Condition 1, the grayscale version of the original image was used. All other conditions were based on this version. In Condition 2 (global luminance reduction), the luminance of each pixel was reduced to half its original value. In Condition 3 (global contrast reduction), the mean image luminance was subtracted from the image, the result divided by 2, and the mean added again. This procedure results in a 50% reduction of image contrast, keeping the mean luminance unchanged. In Condition 4 (local luminance equalization), we first computed mean luminance of the original image in each local region weighted with a Gaussian distribution of 8 pixels of standard deviation in each direction. The original image was then divided point-wise by this local mean. Finally, the result was multiplied by the original image mean and linearly rescaled to maximal displayable luminance range (0.11 to 66 cd/m^2) without changing the mean luminance. This results in a stimulus that has the same mean luminance as the original but no variation of luminance on scales substantially larger than the 16-pixel (0.5°) local region. In Condition 5 (local contrast equalization), local luminance contrast in an 8-pixel-wide region (see also the feature definitions earlier) was computed. As in Condition 3,

the mean luminance was subtracted from the image, but rather than applying a global scaling with 0.5, each pixel of this mean-free image was multiplied with $2<C>/(C(x, y) + <C>)$, where $C(x, y)$ is the contrast around (x, y) and $<C>$ the mean contrast over the image. To the resulting image the mean was added, and—as in Condition 4—the image was linearly rescaled to the maximal luminance range possible without affecting the mean. This procedure results in an image where local contrast is softly equalized: If original contrast is small, contrast is doubled; if local contrast corresponds to the global average, no modification occurs; and if local contrast is large, contrast is reduced. In Condition 6 (random phase), the image was transformed into Fourier space, with the amplitude spectrum preserved but the phase chosen at random. To preserve the large-scale luminance distribution of the image (e.g., bright upper part, dark lower part), the lowest spatial-frequency component in each dimension was not subjected to phase randomization. That is, at points of the spectrum for which either k_x or k_y is below 1 cycle/image, the phase remained intact. Transforming back to image space, this results in a stimulus whose second-order statistics and very-large-scale (1 cycle/image) luminance distribution match the original image but which is deprived of any local high-level structure.

Appendix C: Model definitions

The main models, as used for Figures 5 through 8, are described by the equation

$$y_{si} = \beta_0 + \mathrm{b}_{s0} + \mathrm{b}_{i0} + \sum_{k=1}^{5} (\beta_k + \mathrm{b}_{sk} + \mathrm{b}_{ik}) x_{ksi} + e_{si},$$

where y denotes the response variable (fixation duration), β denotes fixed effects, b denotes random effects, e denotes the residual, the index s marks subjects ($1 \le s \le N_{\mathrm{sub}}$), the index i marks images/items ($1 \le i \le N_{\mathrm{items}}$), and the index k identifies the intercept and the predictors: $k = 0$, the intercept; $k = 1$, the empirical-salience predictor; $k = 2$, the luminance predictor; $k = 3$, the luminance-contrast predictor; $k = 4$, the edge-density predictor; and $k = 5$, the eccentricity predictor.

To compromise between a maximal random-effect structure and computational feasibility, the variance-covariance matrices for the random effects are constrained as follows:

The variance-covariance matrix for by-subject random effects is

$$\Phi_s = \begin{pmatrix} \mathrm{var}(b_{s0}) & \mathrm{cov}(b_{s0}, b_{s1}) & 0 & 0 & 0 & 0 \\ \mathrm{cov}(b_{s0}, b_{s1}) & \mathrm{var}(b_{s1}) & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathrm{var}(b_{s2}) & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathrm{var}(b_{s3}) & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathrm{var}(b_{s4}) & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathrm{var}(b_{s5}) \end{pmatrix}$$

and the variance-covariance matrix for by-item random effects is

$$\Phi_i = \begin{pmatrix} \mathrm{var}(b_{i0}) & \mathrm{cov}(b_{i0}, b_{i1}) & 0 & 0 & 0 & 0 \\ \mathrm{cov}(b_{i0}, b_{i1}) & \mathrm{var}(b_{i1}) & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathrm{var}(b_{i2}) & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathrm{var}(b_{i3}) & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathrm{var}(b_{i4}) & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathrm{var}(b_{i5}) \end{pmatrix}.$$

In both of these matrices, cov() denotes the covariance matrix and var() the variance. That is, only the covariance matrix pertaining to the empirical-map predictor is included explicitly, while the remaining by-item and by-subject covariance matrices are constrained to 0.

For the simplified models of Figure 4 the model equation reduces to

$$y_{si} = \beta_0 + \mathrm{b}_{s0} + \mathrm{b}_{i0} + (\beta_1 + \mathrm{b}_{s1} + \mathrm{b}_{i1}) x_{si} + e_{si}$$

with the corresponding reduced variance-covariance matrices

$$\Phi_s = \begin{pmatrix} \mathrm{var}(b_{s0}) & \mathrm{cov}(b_{s0}, b_{s1}) \\ \mathrm{cov}(b_{s0}, b_{s1}) & \mathrm{var}(b_{s1}) \end{pmatrix}$$

and

$$\Phi_i = \begin{pmatrix} \mathrm{var}(b_{i0}) & \mathrm{cov}(b_{i0}, b_{i1}) \\ \mathrm{cov}(b_{i0}, b_{i1}) & \mathrm{var}(b_{i1}) \end{pmatrix}.$$

In R notation, the simple model is represented as

```
fixDur ~ empMap + (1 + empMap|subNum)
  + (1 + empMap|imgNum)
```

and the main model as

```
fixDur ~ empMap + lum + con + ed + ecc
  + (1 + empMap|subNum)
  + (1 + empMap|imgNum) + (0 + lum|subNum)
  + (0 + lum|imgNum) + (0 + con|subNum)
  + (0 + con|imgNum) + (0 + ed|subNum)
  + (0 + ed|imgNum) + (0 + ecc|subNum)
  + (0 + ecc|imgNum).
```

The variables of the R notation are vectors that each contain one entry per observation, where fixDur refers to fixation durations (y_{si}), empMap refers to empirical salience values ($x_{1\mathrm{si}}$), lum refers to luminance values ($x_{2\mathrm{si}}$), con refers to luminance-contrast values ($x_{3\mathrm{si}}$), ed refers to edge-density values ($x_{4\mathrm{si}}$), ecc refers to eccentricity values ($x_{5\mathrm{si}}$), subNum is the identifier of the subject s, and imgNum is the identifier of the image/item i.